EXCLI Journal 2011;10:240-245 – ISSN 1611-2156
Received: November 11, 2011, accepted: November 22, 2011, published: November 24, 2011

Original article:

PYBACT: AN ALGORITHM FOR BACTERIAL IDENTIFICATION

Chanin Nantasenamat[1,2,*], Likit Preeyanon[1,2], Chartchalerm Isarankura-Na-Ayudhya[2], Virapong Prachayasittikul[2]

[1] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[2] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
* corresponding author: E-mail: mtcnt@mahidol.ac.th, Tel: +662 441 4371, Fax: +662 441 4380

ABSTRACT

PyBact is a software written in Python for bacterial identification. The code simulates the predefined behavior of bacterial species by generating a simulated data set based on the frequency table of biochemical tests from diagnostic microbiology textbook. The generated data was used for predictive model construction by machine learning approaches and results indicated that the classifiers could accurately predict its respective bacterial class with accuracy in excess of 99 %.

Keywords: PyBact; Python; bacteria; bacterial identification; microbiology; data mining

INTRODUCTION

The span of characteristics that can be used for bacterial identification includes the following properties: cultural, morphological, physiological, biochemical, nutritional, chemotaxonomic, serological, inhibitory tests, genotypic properties, chromatographic properties and electrophoretic properties (Logan, 1994; Giacomini et al., 2000). Among these sets, biochemical tests have remained the typical first step for bacterial identification. Biochemical test measures the ability of an unknown microorganism in metabolizing substrates (e.g. sugar, amino acids, etc.). Such catalysis gives rise to detectable color change that is the basis for bacterial identification. Bacterial identification through the use of biochemical tests can be performed by either manual or automated approaches. The former requires analytical skills of technologists in reading and interpreting the biochemical results while the latter makes use of computers for bacterial identification (Sneath, 1964). Such task is typically based on the calculation of probabilities via Bayes' theorem (Lapage et al., 1973; Willcox et al., 1973). This approach essentially generates a quantitative value, known as Willcox's identification scores (Willcox et al., 1980), which measures the similarity of an unknown isolate with those in the data matrix.

Although great advancement has been made for bacterial identification methods, there still exist unforseeable factors that may potentially hinder accurate and definitive bacterial identification (Frederiksen and Tønning, 2001). Suggestions for preventing and resolving such potential misidentification have been proposed (Janda and Abbott, 2002). Therefore, there is still ample room for improving upon the methodologies for bacterial identification.

This study describes a simple algorithm implemented using the Python programming language for bacterial identification. Python is potentially suited for such appli-

EXCLI Journal 2011;10:240-245 – ISSN 1611-2156
Received: November 11, 2011, accepted: November 22, 2011, published: November 24, 2011

cation as it is an open source, multiplatform, easy to learn and implement, has extensive modules and libraries as well as having a collaborative online community (Biopython Project, 2008). Furthermore, the Python programming language has been extensively demonstrated to be useful for many endeavors especially those in the life sciences (Bassi, 2007; Python, 2008).

MATERIAL AND METHODS

Compilation of data set

The data sets used in this study is comprised of 12 species from *Vibrio* genus and 134 species from *Enterobacteriaceae* family. The positive percentage of biochemical profile of these bacteria was obtained from diagnostic microbiology textbook (Murray et al., 2007) and used as input data for further processing by PyBact. (Note: the original data set of these biochemical profiles of *Vibrio* and *Enterobacteriaceae* is provided as a supplementary information).

Algorithm for biochemical data generation

PyBact is freely available under the Open Software License (OSL) Version 3.0 at http://pybact.sourceforge.net/. Up until now, there has been more than 350 downloads worldwide. The principle behind the software is based on the generation of simulated binary data, which is essentially a table of strains versus biochemical tests. The generated data matrix serves as processable input data for machine learning algorithms, which could be applied for improved and robust bacterial identification purposes. Aside from this, the software holds great benefit as an educational aid for diagnostic microbiology courses.

The algorithm of PyBact's data generation procedures is summarized as a pseudocode in Table 1 and the generated data matrix is represented in Figure 1. Briefly, each biochemical test is represented by a binary value of either 1 or 0, which corresponds to positive or negative results of the biochemical property of interest.

The algorithm initially constructs *N* number of strains for a given bacterial species of interest. For each strain, the biochemical tests of interest were then randomly assigned a value of either 1 or 0; the value of 1 is assigned until the cumulative count of 1 equals the predefined positive percentage. For example, if the positive percentage is 95 %, the value of 1 will be randomly assigned until the cumulative count of 1 equals 95.

Table 1: PyBact's data generation algorithm

```
import modules: random, os, sys, and psyco
read data from input file
construct an array and import input data as array members
for each bacteria (from a list of bacterial species):
    generate class label for N number of strains
    for each biochemical test (from a list of biochemical tests):
        populate the value as 0 (representing negative results):
        if the biochemical test has a predefined positive percentage of 0
        then populate all value as 0
        else
            for each biochemical test (from a list of biochemical tests):
                assign a value of 1 (representing positive results) to a random position
                until the cumulative count of 1 equals the predefined positive percentage
            append biochemical test to the list
write results to output file in the form of a tab delimited data matrix
```

EXCLI Journal 2011;10:240-245 – ISSN 1611-2156
Received: November 11, 2011, accepted: November 22, 2011, published: November 24, 2011

The input data matrix is depicted as follows (showing only the first 5 biochemical test and the first 2 bacterial species):

Indole	MR	VP	Citrate	H_2S	
99	99	75	97	0	*V.cholerae*
98	99	9	99	0	*V.mimicus*

After the algorithm's procedures have been performed, the generated output yields the following data matrix (this example shows only the first 5 biochemical test and 5 strains of the first 2 bacterial species):

Indole	MR	VP	Citrate	H_2S	
1	1	0	1	0	*V.cholerae*
0	1	0	1	0	*V.cholerae*
1	0	1	1	0	*V.cholerae*
1	1	1	0	0	*V.cholerae*
1	1	1	1	0	*V.cholerae*
1	1	0	1	0	*V.mimicus*
1	1	0	0	0	*V.mimicus*
1	1	0	1	0	*V.mimicus*
0	0	0	1	0	*V.mimicus*
1	1	0	1	0	*V.mimicus*

The dynamic nature of the algorithm makes certain that the generated data matrix is unique with each run. This feature resembles the dynamic nature of how each bacterial strain of a certain genus and species could possess slight variations in the biochemical profiles.

Generating the biochemical data

As previously mentioned, PyBact is available at http://pybact.sourceforge.net/ as a zip file named PyBact_1.0.1.zip, the contents of which contains the source code, the compiled program, the input files for *Vibrio* and *Enterobacteriaceae* data set and a readme text file. The input file is essentially a frequency table of positive occurrence of biochemical test for bacterial species as obtained from the Manual of Clinical Microbiology (Murray et al., 2007). To generate the simulated biochemical data with PyBact, the following command is entered into the command prompt of Windows:

```
pybact input.txt output.txt 100
```

where input.txt and output.txt are, as the name implies, the input and output files. 100 is the number of isolates to be generated per species. The output file is tab delimited and can readily be used as data set for machine learning analysis. The program is flexible in which the number of biochemical tests it can handle are automatically determined from the input file. In addition, the number of isolates can be adjusted to produce small or large data sets as desired.

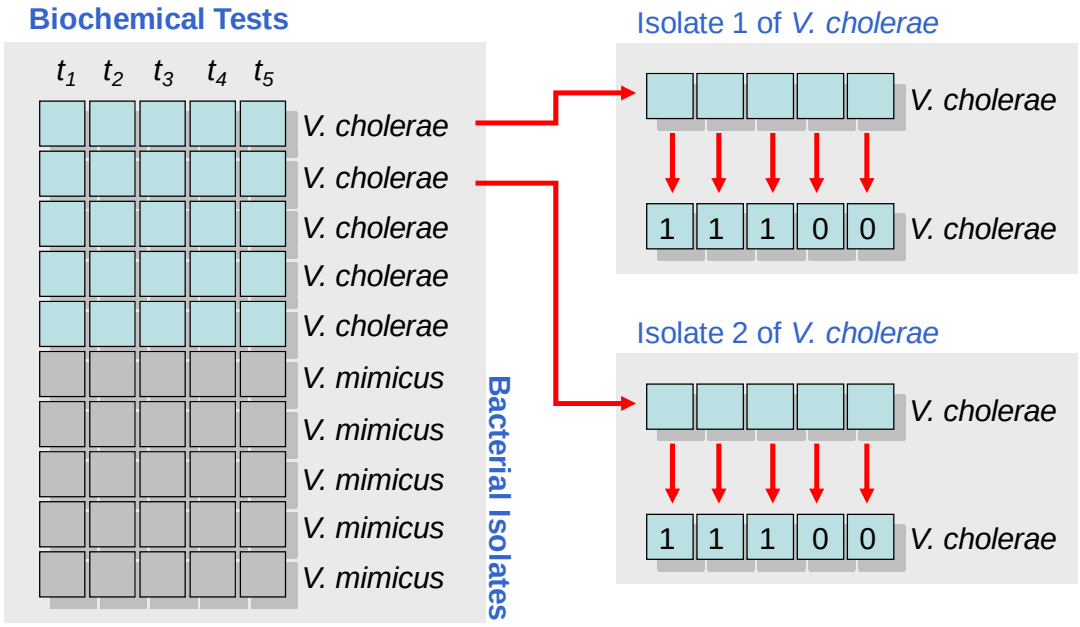

Figure 1: Schematic representation of biochemical test data generation by PyBact

EXCLI Journal 2011;10:240-245 – ISSN 1611-2156
Received: November 11, 2011, accepted: November 22, 2011, published: November 24, 2011

Data mining for bacterial identification

Data mining is a technique that seeks to find correlation between a set of independent variables (e.g. the biochemical test profiles) with the observed dependent variable (e.g. the bacterial species). The principles of using data mining techniques on biological and chemical data sets have previously been reviewed (Nantasenamat et. al., 2009, 2010). Successful utilization of data mining techniques for modeling problems in biology and chemistry have previously been demonstrated (Nantasenamat et al., 2005, 2007a, b, 2008a, b; Thippakorn et al., 2009; Worachartcheewan et al., 2009, 2010a, b, 2011).

Bacterial identification was performed using Weka (Witten and Frank, 2005) and employing the generated data matrix of biochemical test data, which was obtained from PyBact, as the input data. The classifier used for bacterial identification included decision tree, naïve Bayes and support vector machine. The classification was performed using default parameter and 10-fold cross-validation. A schematic representation of the methodology used in this study is depicted in Figure 2.

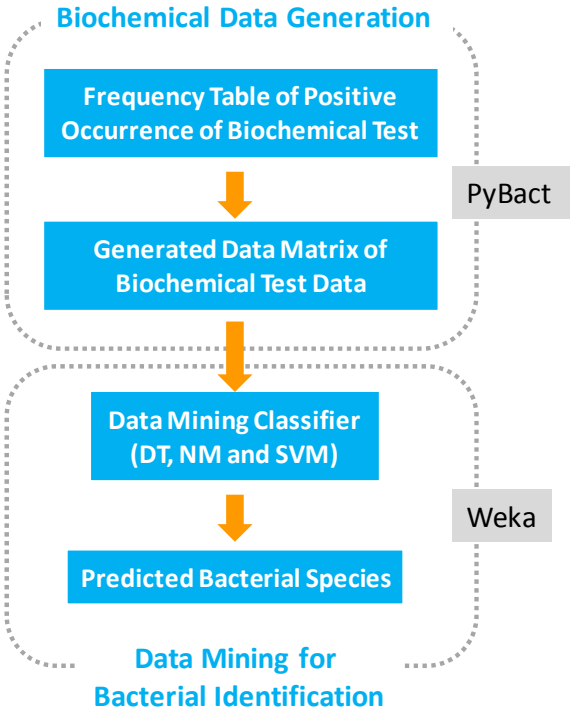

Figure 2: Schematic representation of bacterial identification methodology

RESULTS AND DISCUSSION

To demonstrate the applicability of the generated binary data matrix for bacterial identification, the generated output was then used as input data for bacterial identification by machine learning classifiers of Weka. The data set was tested with the following classifiers using 10-fold cross-validation: decision tree, naïve Bayes and support vector machine.

The data set used for this analysis is made up of 12 species from *Vibrio* genus. The number of strains to generate for each of the species was set to 100. Therefore, the data set is comprised of a total of 1,200 *Vibrio* strains (12 species from *Vibrio* genus × 100 strains).

Bacterial identification was performed using the aforementioned classifers with default parameter. Results indicated that the classifiers could accurately identify the correct bacterial species by using the biochemical test data for pattern recognition. It was observed that the decision tree implementing the J48 algorithm could accurately identify 1,189 out of 1,200 strains which equates to 99.0833 %. Furthermore, naïve Bayes and support vector machine could correctly identify all of the 1,200 strains to give 100 % accuracy.

The aforementioned results indicated that the generated data set in combination with machine learning algorithms is applicable for differentiating the various species of *Vibrio*. Question arises as to whether the approach described herein is also applicable for bacterial identification of mixed populations, particularly, the ability in differentiating various genus and species. To answer such question, our program was applied to the largest and most heterogeneous family of medically important bacteria, the *Enterobacteriaceae*. The data set used for this analysis was comprised of 134 species. As stated earlier, the number of strains to generate for each of the species was set to 100. Therefore, the data set is comprised of a total of 13,400 strains (134 species from

Enterobacteriaceae family × 100 strains). The data set was also tested with decision tree, naïve Bayes and support vector machine which could accurately identify 12,297 (91.7687 %), 12,701 (94.7836 %) and 12,691 (94.7090 %) species, respectively.

CONCLUSION

We have proposed a novel approach for employing Python in generating simulated data set for constructing a predictive model useful in bacterial identification. The flexibility of the algorithm lies in its ability to mimic the dynamic nature of positive/negative occurrence probability of the biochemical test results. Moreover, the generated binary data matrix has been demonstrated to be useful in machine learning analysis as observed by its ability to correctly predict the bacterial species of all isolates. The proposed algorithm and data set could easily be adapted and modified to build robust tools for accurate and precise bacterial identification that are of research, professional and educational value.

ACKNOWLEDGEMENTS

This work was supported in part by the Goal-Oriented Research Grant from Mahidol University and the Science and Technology Research Grant from the Thailand Toray Science Foundation. Partial support is also acknowledged from Office of the Higher Education Commission and Mahidol University under the National Research Universities Initiative.

REFERENCES

Bassi S. A primer on python for life science researchers. PLoS Comput Biol 2007;3: e199.

Biopython, http://www.biopython.org.

Frederiksen W, Tønning B. Possible misidentification of *Haemophilus aphrophilus* as *Pasteurella gallinarum*. Clin Infect Dis 2001;32:987-9.

Giacomini M, Ruggieroa C, Calegari L, Bertone S. Artificial neural network based identification of environmental bacteria by gas-chromatographic and electrophoretic data. J Microbiol Meth 2000;43:45-54.

Janda JM, Abbott SL. Bacterial identification for publication: when is enough enough? J Clin Microbiol 2002;40:1887-91.

Lapage SP, Bascomb S, Willcox WR, Curtis MA. Identification of bacteria by computer: general aspects and perspectives. J Gen Microbiol 1973;77:273-90.

Logan NA. Bacterial systematics. Oxford: Blackwell Scientific Publ., 1994.

Murray PR, Baron EJ, Jorgensen JH, Landry ML, Pfaller MA. Manual of clinical microbiology. Washington DC: ASM Press, 2007.

Nantasenamat C, Naenna T, Isarankura Na-Ayudhya C, Prachayasittikul V. Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network. J Comput Aid Mol Des 2005;19:509-24.

Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Quantitative structure-imprinting factor relationship of molecularly imprinted polymers. Biosens Bioelectron 2007a;22:3309-17.

Nantasenamat C, Isarankura-Na-Ayudhya C, Tansila N, Naenna T, Prachayasittikul V. Prediction of GFP spectral properties using artificial neural network. J Comput Chem 2007b;28:1275-89.

Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Prediction of bond dissociation enthalpy of antioxidant phenols by support vector machine. J Mol Graph Model 2008a;27:188-96.

Nantasenamat C, Piacham T, Tantimongcolwat T, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR model of the quorum-quenching N-acyl-homoserine lactone lactonase activity. J Biol Syst 2008b;16:279-93.

Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A practical overview of quantitative structure-activity relationship. EXCLI J 2009;8:74-88.

Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Advances in computational methods to predict the biological activity of compounds. Exp Opin Drug Discov 2010;5:633-54.

Python, Application Domains. http://www.python.org /about/apps.

Sneath PH. New approaches to bacterial taxonomy: use of computers. Annu Rev Microbiol 1964;18:335-46.

Thippakorn C, Suksrichavalit T, Nantasenamat C, Tantimongcolwat T, Isarankura-Na-Ayudhya C, Naenna T et al. Modeling the LPS neutralization activity of anti-endotoxins. Molecules 2009;14:1869-88.

Willcox WR, Lapage SP, Bascomb S, Curtis MA. Identification of bacteria by computer: theory and programming. J Gen Microbiol 1973;77:317-30.

Willcox WR, Lapage SP, Holmes B. A review of numerical methods in bacterial identification. Antonie Van Leeuwenhoek 1980;46:233-99.

Witten IH, Frank E. Data mining: practical machine learning tools and techniques. San Francisco: Morgan Kaufmann, 2005.

Worachartcheewan A, Nantasenamat C, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Modeling the activity of furin inhibitors using artificial neural network. Eur J Med Chem 2009;44:1664-73.

Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V. Identification of metabolic syndrome using decision tree analysis. Diabetes Res Clin Pr 2010a;90:e15-8.

Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V. Lower BMI cutoff for assessing the prevalence of metabolic syndrome in Thai population. Acta Diabetol 2010b;47:S91-6.

Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul S, Prachayasittikul V. Predicting the free radical scavenging activity of curcumin derivatives. Chemometr Intell Lab Syst 2011; 109:207-16.